Multimodal Person Identification through the Fusion of Face and Voice Biometrics

Marwa Saleh, Advisor: Ismail Joumy
Department of Electrical and Computer Engineering, Lafayette College, Easton, PA 18042, USA

Introduction

- Person identification is used every day in a variety of applications from access control and security cameras to social media and face ID on smartphones.
- The process of person identification involves different biometrics such as face, voice, fingerprint, etc.
- However, there are still various limitations to their efficiency which makes them incompletely reliable.
- In public settings, like that of a museum, using one biometric alone becomes challenging due to factors like background noise, overlap of people’s faces, varying angles and/or distances from the camera, as well as the recently introduced challenge of face masks.

Feature Extraction

Viola-Jones Algorithm:
1. Divide image into squares
2. Calculate delta of sum (shaded) and sum (unshaded)
3. Identify any Haar-like features
4. Crop the 227 x 227 px square where a face was identified

Pitch:

Mel Frequency Cepstral Coefficients (MFCCs):

Conclusions & Future Directions

Results prove that biometric fusion improves the accuracy of person identification compared to using a single biometric.

Future work will include:
- Expanding the dataset to include more subjects
- Testing other convolutional neural networks
- Testing other fusion strategies (e.g., feature-level)
- Using greyscale edge detection for a color-blind face recognition system
- Implementing a two-stage system that first passes through a binary classifier in order to minimize the number of possible classes in the second stage

References


Acknowledgments

David and Lorraine Freed Undergraduate Research Symposium, Lehigh University

Methodology

Importing Dataset
Preprocessing
Extract frames from each video
Convert all stereo audio to mono audio
Remove all silence and unvoiced speech

Feature Extraction

Implement Viola-Jones algorithm to detect faces
Extract Pitch and MFCC Coefficients

Classification using CNNs

Decision-Level Fusion

Dataset

Michigan State University Audio-Video Indoor Surveillance (MSU-AVIS) Dataset:
- 50 Subjects (16 females, 34 males)
- Image data variations include:
  - Indoor illumination
  - Facial expressions
  - Pose & distance relative to the camera
- Audio data variations include:
  - Indoor reverberations
  - Background Noise
  - Distance from the microphone

Fusion

Three decision-level fusion algorithms were implemented:
1. Fusing by higher confidence
2. Fusing by higher confidence after normalization
3. Fusing by higher entropy of confidence scores